2011

Proficiency Tests to Estimate Error Rates in the Forensic Sciences

Jonathan Koehler
Northwestern University School of Law, jay.koehler@northwestern.edu

Proficiency Tests to Estimate Error Rates in the Forensic Sciences

Jonathan J. Koehler, Ph.D.
Northwestern University School of Law
9-19-10

A proficiency test is an assessment of the performance of laboratory personnel using samples whose sources are known to the proficiency test administrator but unknown to the examinee. There are many purposes to such assessments: training personnel, ensuring that personnel achieve baseline competence levels, improving laboratory practices and procedures, and identifying future needs for a laboratory. Proficiency tests can also help identify *reasonable first pass estimates for the rates at which various types of errors occur*.

It is crucial to obtain error rate estimates because the reliability and probative value of forensic science evidence is inextricably linked to the rates at which examiners make errors. Without such information, legal decision makers have no scientifically meaningful way of thinking about the risk of false identification and false non-identification associated with forensic reports.

When designing proficiency tests to estimate error rates, careful thought must be given to at least four issues:
(1) the composition of the test designers and administrators who oversee the testing process,
(2) the features of test and reference samples,
(3) the composition and selection of test participants, and
(4) the use of blind test protocols.

Issue #1: Test Designers and Administrators

The designers and administrators of proficiency tests should be *qualified, disinterested parties.* By "qualified," I mean people who have expertise in such areas as experimental design, testing, statistics, behavioral sciences, police investigation, and forensic science. It would be hard to overstate the importance of including statisticians, behavioral scientists and others who have training and experience in matters related to research methodology involved in this process. If the proficiency tests are not properly designed, then scientific inferences cannot be made. By "disinterested," I mean that proficiency test designers and administrators should not be affiliated with the examinees or the examinees' laboratories, nor should they stand to benefit from or be harmed by any particular outcome or set of outcomes on the proficiency tests.

Issue #2: Features of Test Samples

The samples used in proficiency tests should approximate a random sample of the types of evidentiary samples that arise in actual cases. This may be accomplished in different ways. One way is for test administrators to access a database of all cases in a county, state, country, or other population over some time period (e.g., five years), and to note which cases included forensic science evidence. A random sample of those cases might then be identified as prototypes for the manufacture of proficiency test samples. Samples identified in this manner are likely to vary widely. Using a fingerprint example: one case might include two detailed latents plus rolled prints from one suspect and two innocents. Another case might include one badly smudged latent and rolled prints from each of ten suspects, including a pair of identical twins.

Once a random sample of cases has been identified, test administrators should write comparable cases and then manufacture forensic evidence that resembles the samples and cases chosen. Materials should not be reused across tests.

The newly created evidence should be rated for difficulty using an agreed-upon rating scheme to ensure that they parallel the sample of selected cases and to allow researchers to track the impact of sample difficulty on examiner accuracy. Likewise, administrators should track task features such as whether multiple samples are from a single common source, and whether the source of the print or marking is or is not present.

Issue #3: Test Participants

Test participants should be a random or otherwise representative sample of forensic scientists who testify in court. Pertinent background features of selected participants should be tracked. These features should include training, experience, and number of cases in which participants have testified. By tracking examiner characteristics, we will gain insight into the conditions under which performance varies.

All forensic scientists who testify in court must be part of the participant pool. Examiners cannot opt in or out. However, it is not important that all or even most forensic scientists be selected to participate in the proficiency tests. The idea is that participants are sampled using statistically sound methods. This method will allow for extrapolation of results to the broader forensic population.

The notion that forensic scientists should be selected at random rather than required to participate in any given test may come as a surprise. But it is consistent with the proficiency testing purpose described here: to *identify a reasonable first pass estimate for the rates at which various types of errors occur.* The purpose of the tests is not to identify lab-specific or examiner-specific error rates. Such data would be useful, but they are difficult to obtain, and likely to be misinterpreted or dismissed even if they were obtained.[1]

The notion that our focus should be on identifying general error rates rather than individual or situation-specific error rates is worth careful consideration because intuition suggests that the opposite is true. After all, why should a careful, well-trained, and experienced examiner be saddled with the same first pass error rate estimate as a careless, poorly trained, and inexperienced examiner?

The answer to this loaded question lies in the name of the error rate itself. It is a "first pass" error rate estimate. It is an estimated base rate for errors. Factfinders need such base rates, in combination with individuating information about a given examination and a given examiner, to make an informed judgment about the risk of error in any particular case. The job of the proficiency tests described here is to provide that base rate. It is not to provide the individuating information that the factfinder might use to adjust the base rate.

A sports analogy provides some clarification. In order to know the chance that a professional baseball player will get a hit in his next at bat, a forecaster needs to have a sense of the base rate

[1] Some people favor proficiency tests that identify examiner-specific error rates rather than general error rates. Although such data are desirable in the abstract, they are unlikely to be helpful in practice. In any endeavor where errors are infrequent, too many tests are required under too many different conditions to identify person-specific error rates that apply to particular cases. Consider, for example, an examiner who makes 100 comparisons without error under controlled, rigorous and realistic test conditions. What does this test performance tell us about the examiner's error rate? From a statistical standpoint, zero errors out of 100 trials represents an underlying error rate anywhere from 0% (the examiner theoretically could be "perfect") to about 3%. And that 3% figure (which is the approximate upperbound of a 95% confidence interval) would be the cause of a lot of anger and misunderstanding ("How could the examiner have a 3% error rate when he/she didn't make any errors?"). The confusion can be resolved by a class in statistical inference. But judges, jurors and experts who lack statistical training will not appreciate this point. And even those who do appreciate the inferential uncertainty surrounding test performance are likely to dismiss the entire effort on grounds that the conditions in the focal case differ in significant ways from those in the test conditions. In the end, then, even if it were possible to get busy, backlogged examiner to sit for a lengthy set of proficiency tests designed to estimate individual rates of error, the data would cause confusion and prejudice at trial.

for hits. Baseball players get hits about one time in four chances. A good player gets a hit one time in 3.5 chances, and an outstanding player gets a hit one time in three chances. These are base rates. If individuating information is available – such as whether the player has an injury, is facing a tough pitcher, or is hitting with the wind blowing out – adjustments to the base rate should be made. But the base rate is the anchor. The base rate for a hit in baseball is very different than the base rate for a goal in professional hockey (about 10%) or for a successful free throw in professional basketball (about 75%). These first pass sports base rates are known. In contrast, the base rates for errors in the forensic sciences are unknown and cannot be discerned from case reports or legal outcomes: scientific study is needed.

Issue #4: Blind Tests

Ideally, proficiency tests should be blind in the sense that any party that has a direct interest in how the examiners perform should not be aware that the proficiency test materials are part of a test rather than part of actual casework. Behavior may change under observation and it is important to make test conditions as similar to casework conditions as possible. Part of that similarity means not telling examiners that they are being tested. This is a key feature in a scientifically valid proficiency test of human performance.

Some may take offense at the suggestion that forensic scientists' behavior may vary when they know they are being tested. After all, they are trained professionals and many have years of experience. But the notion that behavior changes under observation is well-documented across many domains for both experts and novices alike. It is simply part of the human condition.

The notion of blindness in proficiency tests is sometimes dismissed on grounds that it either cannot be done or would be too costly to implement. These criticisms should be rejected. Blind proficiency testing is already used in some forensic science areas (such as the DOD's forensic urine drug testing program and HIV testing program). Blind tests have also been used for DNA analyses. For example, a 2002 study in the *Int'l J. of Legal Medicine* reports the results of DNA blind trials across 129 laboratories in 28 European countries. Furthermore, Joe Peterson has conducted a detailed pilot investigation in the U.S. which showed that blind testing of DNA analysts can be done. His reports appear in two articles in the *Journal of Forensic Sciences* in 2003 and in detailed discussions in papers filed with the National Institute of Justice.

Conclusion

As noted above, the reliability and probative value of forensic science evidence is inextricably linked to the rates at which examiners make errors. Jurors and others cannot assess the significance of a reported forensic science match without having some sense of the rate at which false positive errors occur. Properly designed proficiency tests provide a necessary first step. The design and administration of these proficiency tests – i.e., tests that will provide reasonable first pass estimates for the rates at which various types of errors occur – is a major undertaking. And even a successful venture that identifies error rate estimates for different technologies under various conditions will not tell us everything we wish to know about the risk of error in specific cases. But if they provide jurors and others with a better sense of the probative value of a reported match, they will serve their purpose.

Of course, it would be easy to dismiss the entire enterprise described here by pointing out unanswered practical questions. Who will administer the tests? How much will they cost? How can participation be ensured? How will examiners create enough time to participate? These are important questions. But they are not the questions that the forensic science community in general, and the fingerprint community in particular, should be asking at this stage. Instead, the $64,000 question is this: *is the forensic science community prepared to accept the idea that error rates matter, and that the way to estimate those rates is through a carefully designed, rigorously scientific, testing program?* If the answer is "yes," then we are half way there.